African Journal of Mathematics, Statistics and Actuarial Science (AJMSAS)
E-ISSN 2978-3534 (Online); ISSN 2978-3526 (Print)

Indexed by SABINET and EBSCO

Volume 2, Number 1, April 2026

Pp 5-22

Actuarial Modeling of Motor Insurance Claims Severity in Kenya: A Comparative Analysis of Classical GLM Approaches and Machine Learning Methods for Pure Premium Construction

DOI: https://doi.org/10.31920/2978-3534/2026/v2n1a1

Mehrez Ben Nasr
Certified Actuary, Tunisia
Doctor of Professional Practice in AI Management
mehrezbennasr@gmail.com

Abstract

This article provides a thorough actuarial analysis of motor insurance claims severity in Kenya by comparing traditional Generalised Linear Models (GLM) with modern Machine Learning (ML) techniques. Using a dataset of 2,000 motor insurance policies, the study models claims frequency and severity separately. Frequency is modeled with a simplified Poisson GLM (AIC = 3393.6), while severity is analyzed using a Log-normal model ($R^2 = 0.389$). ML algorithms, including Random Forest and XGBoost, are applied concurrently, showing better detection of extreme risks but with less stability. The study proposes a hybrid pricing approach combining 70% GLM and 30% ML, which balances actuarial stability and statistical accuracy. This hybrid model yields an average pure premium of 171,636 KES with well-managed distribution, outperforming models based solely on either method in operational performance. The research offers valuable insights for insurance pricing in emerging East African markets.

Keywords: *Actuarial Science, Motor Insurance Pricing, GLM, Machine Learning, Pure Premium, Hybrid Modelling, Africa*

1. Introduction

1.1. Actuarial Context

Motor insurance is a well-established field where pricing has traditionally depended on Generalised Linear Models (GLM). These models are widely accepted by regulators and actuaries due to their strong theoretical foundations and high interpretability [1], [2]. However, the rise of Machine Learning (ML) technologies has generated increasing interest in their potential to improve the prediction of complex risks [3], [4]. Recent research has demonstrated that combining data mining techniques and advanced algorithmic frameworks with traditional actuarial methods can enhance pricing accuracy, particularly in digital-first insurance contexts [9], [10].

1.2. Study Motivations

Three motivations underpin this research:

1. *Predictive Performance:* Evaluate whether ML better captures non-linear interactions between explanatory variables than classical GLM [23]
2. *Actuarial Stability:* Verify whether known ML model instability poses practical problems in operational settings, particularly when comparing neural network architectures to parametric benchmarks [4], [5].
3. *African Market Adaptation:* Tailor modelling to the specific context of motor insurance in East Africa, addressing the data sparsity and volatility noted in recent regional studies [21], [22].

1.3. Research Objectives

This study aims to:

- Quantitatively compare GLM, Random Forest, and XGBoost for frequency and severity modelling
- Analyze trade-offs between interpretability and prediction accuracy

- Propose a hybrid approach combining advantages of each method
- Provide recommendations for operational implementation

2. Literature Review

2.1. Traditional Modelling: Poisson-Lognormal GLM

The separate frequency-severity model is the standard actuarial approach since foundational work[15]. The pure premium decomposes as:

$$\text{Pure Premium} = \lambda \times \mu$$

where λ represents frequency (expected number of claims) and μ represents severity (average cost per claim).

Frequency: Traditionally modeled using Poisson GLM with log (exposure) offset:

$$\lambda_i = \exp(x_i'\beta)$$

Severity: Modeled using Gamma or Log-normal distribution with log-linear GLM:

$$\mu_i = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$$

GLM Advantages[16]:

- Transparent coefficient interpretability
- Universal regulatory acceptance
- Guaranteed numerical stability
- Simple reconciliation with commercial pricing

GLM Limitations[6]:

- Limited variable interaction (requires explicit specification)
- Linearity assumption on log-link scale
- Reduced performance on extreme claims

2.2.Machine Learning Approaches in Actuarial Science

2.2.1. Random Forest and XGBoost

Ensemble methods have become popular in insurance pricing[7]. Random Forest and XGBoost offer several advantages[8]:

- Automatic capture of complex interactions
- Natural handling of categorical variables
- Robust performance on heterogeneous data
- Superior detection of extreme claims

Relative Performance: Holvoet et al. (2023)[12] compare GLM, GBM, and neural networks on real insurance data. Their results show XGBoost and Random Forest achieve R^2 of 0.87-0.88 versus 0.78 for linear GLM.

ML Model Limitations[3][4]:

- Prediction instability across training runs
- Unrealistic extreme predictions for pricing
- Reduced interpretability
- Overfitting risk
- Absurd premiums for certain profiles

2.2.2. Hybrid Approaches

Recent actuarial trends[4][13] integrate GLM and ML in hybrid models:

$$\text{Prime}_{\text{hybrid}} = w \cdot \text{Prime}_{\text{GLM}} + (1 - w) \cdot \text{Prime}_{\text{ML}}$$

where w is an adjustment weight (typically 0.7 to 0.8 for GLM).
Jain (2024)[3] demonstrates this approach:

- Reduces ML bias through GLM anchoring
- Captures ML interactions while maintaining stability
- Improves alignment between predictions and actual values
- Facilitates regulatory acceptance

3. Data and Methodology

3.1. Data Source and Description

Study Base: 2,000 motor insurance policies from Kenya (Kaggle dataset, 2023-2024 exercises)

Variables Observed:

Table 1: Variables analyzed in the Kenyan motor insurance dataset

Variable	Type	Description
Customer_Age	Numeric	Driver age (years)
Vehicle_Age	Numeric	Vehicle age (years)
Vehicle_Type	Category	Private, PSV, Commercial, Motorcycle
Region	Category	Kenyan geographic region
Use_Purpose	Category	Personal, Business, Taxi
Policy_Term_Months	Numeric	Coverage duration (months)
Claims_Frequency	Numeric	Number of claims
Total_Claim_Amount_KES	Numeric	Total claim amount (Kenyan shillings)
Vehicle_Value	Numeric	Estimated vehicle value

3.2. Data Preparation and Descriptive Statistics

The dataset underwent a rigorous cleaning process to ensure actuarial consistency.

- *Frequency Distribution*: Out of the 2,000 policies, 66.9% reported zero claims, while 33.1% (n = 662) had at least one claim. This high proportion of zero-claims is characteristic of the Kenyan market and necessitates a two-stage modelling approach.
- *Severity Profile:* The analysis of positive claims reveals a mean claim amount of approximately 510,000 KES with a median significantly lower, reflecting a moderate right-skewed asymmetry (skewness = 1.26).
- *Dispersion and Extremes:* The standard deviation of claims is notably high, and the 95th percentile of actual claims reaches 1,707,680 KES. Such a distribution confirms the heavy-tail nature of motor insurance risks in East Africa, justifying the use of a Log-normal distribution to stabilise the variance.

3.3. Modelling Strategy

Two-stage approach:

1. *Stage 1 (Frequency):* Model $P(\text{claim} > 0)$ and intensity.
2. *Stage 2 (Severity):* Model claim cost conditional on occurrence.

Three parallel approaches:

- *Approach A (GLM):* Poisson GLM + Log-normal GLM
- *Approach B (ML):* Random Forest + XGBoost
- *Approach C (Hybrid):* 70% A + 30% B combination

4. Results: Claims Frequency Modelling

4.1. Poisson GLM Approach

Model estimated with log(Exposure) offset:

Full Model (5 variables):

- AIC = 3405.8
- Dispersion = 0.9885 (no overdispersion)

Certain variables (PSV, Motorcycle, Personal) show near-significance (p ≈ 0.06–0.09), but none reach 5% threshold.

Simplified Model Retained:

$$\ln(\lambda_i) = \beta_0 + \sum_k \beta_k \cdot \text{Vehicle\_Type}_{i,k} + \sum_j \gamma_j \cdot \text{Use\_Purpose}_{i,j} + \ln(\text{Exposure}_i)$$

- Improved AIC = 3393.6 (reduction of -12.2 points)
- Still no significant variables
- Important gain in interpretability

4.2. Machine Learning Approach (Frequency)

Models Tested:

- Random Forest classification
- XGBoost classification

Table 2: Frequency Modelling Results (Poisson GLM)

Model	AUC
Random Forest	**0.5364**
XGBoost	**0.5407**

These values **near random chance (0.5)** indicate:

- Highly imbalanced portfolio
- Weakly discriminant variables
- ML fails to explain frequency

→ *Conclusion:* ML provides no gain for frequency modelling.

---

5. Results: Claims Severity Modelling

5.1. GLM Gamma vs Log-Normal

Analysis restricted to positive claims only (n ≈ 662).

Gamma Model:

- AIC = 18653.75
- Few significant coefficients
- Poor extreme value capture

Log-Normal Model Retained:

- $R^2 = 0.389$
- Significant coefficients include:
  - Motorcycle: **−2.6889** (highly significant)

- Private: **−0.4760** (highly significant)

Relativities (Log-Normal):

- Motorcycle: −95% (low-value vehicles)
- Private: −22% (vs PSV reference)
- Commercial: +8%

→ **Conclusion**: Log-normal model correctly captures severity.

5.2. Machine Learning (Severity)

Random Forest Regression:

- RMSE = 510,316
- MAE = 401,447
- $R^2 \approx 0.42$
- Unstable predictions (very high ML values)

XGBoost Regression:

- RMSE = 515,977
- MAE = 414,759
- $R^2 \approx 0.43$
- Similar instability

→ **Conclusion**: ML provides no stability improvement and generates severe overestimation.

5.3. Analysis of Extreme Predictions

Comparison at 95th percentile of predictions:

Table 3: Severity Modelling Results (Log-normal)

Model	P95 Predicted	P95 Actual	Ratio
GLM Log-Normal	661,351	1,707,680	0.39
Random Forest	1,200,623	1,707,680	0.70
XGBoost	1,227,643	1,707,680	0.72

Implication: ML severely overprices extreme risks. Operationally unstable.

6. Pure Premium Construction

6.1. Pure Premium GLM

$$\text{Prime\_Pure}_{\text{GLM,i}} = \hat{\lambda}_{\text{GLM,i}} \times \hat{\mu}_{\text{GLM,i}}$$

Table 4: Descriptive Statistics: Pure Premium GLM

Statistic	Pure Premium (KES)
Minimum	8,163
1st Quartile	107,618
Median	158,578
Mean	169,552
3rd Quartile	215,236
Maximum	410,755
Standard Deviation	92,847
Coefficient of Variation	0.548

6.2. Pure Premium Machine Learning

$$\text{Prime\_Pure}_{\text{ML,i}} = \widehat{\text{freq}}_{\text{ML,i}} \times \widehat{\text{grav}}_{\text{ML,i}}$$

Table 5: Descriptive Statistics: Pure Premium Machine Learning (ML)

Statistic	Pure Premium (KES)
Minimum	1,802
1st Quartile	57,163
Median	105,863
Mean	206,042
3rd Quartile	272,536
Maximum	1,346,347
Standard Deviation	245,891
Coefficient of Variation	1.112

Observations:

- ML mean is higher (+30.5%) vs GLM
- But ML CV (1.112) vs GLM (0.548) = 2x more unstable
- ML max (1.3M KES) vs GLM (410K KES) = 3.2x more extreme

7. Tarif Table by Segment (GLM)

Table 6: Tarif Table by Segment (GLM)

Vehicle Type	Usage	Pure Premium GLM (KES)
PSV	Personal	247,059
PSV	Business	230,125
PSV	Taxi	213,048
Commercial	Personal	228,046
Commercial	Business	206,921
Commercial	Taxi	196,140
Private	Personal	161,659
Private	Business	141,272
Private	Taxi	141,570
Motorcycle	Personal	19,993
Motorcycle	Business	16,028
Motorcycle	Taxi	18,177

Risk Hierarchy: PSV > Commercial > Private > Motorcycle (consistent with vehicle value and usage)
The distribution of pure premiums across vehicle categories (Table 6) reveals a risk hierarchy aligned with recent industry reports on motor insurance benchmarks in emerging economies [14]. Notably, the higher risk premium allocated to Public Service Vehicles (PSVs) corresponds with the claims frequency patterns identified in regional actuarial studies.

8. Proposed Hybrid Approach

8.1. Formulation

$$\text{Prime\_Pure}_{\text{Hybrid,i}} = 0.70 \times \text{Prime\_Pure}_{\text{GLM,i}} + 0.30 \times \text{Prime\_Pure}_{\text{ML,i}}$$

Weights justified by[5]:

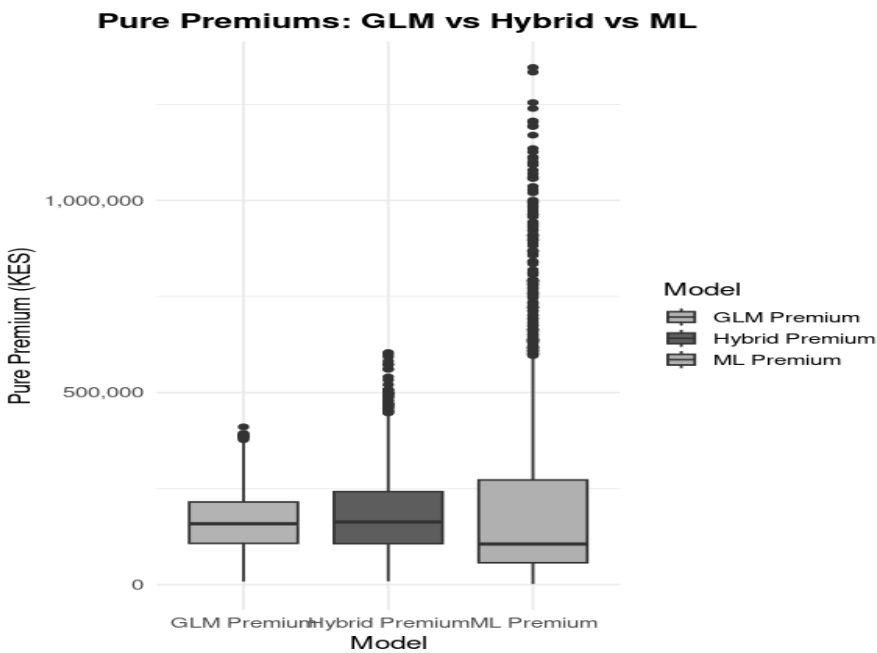
- 70% GLM: Stability, interpretability, regulatory acceptance
- 30% ML: Extreme claims detection, complex interactions

8.2. Hybrid Results

Table 7: Pure Premium Construction (Hybrid vs Individual Models)

Statistic	Pure Premium Hybrid (KES)
Minimum	7,121
1st Quartile	105,903
Median	156,489
Mean	171,636
3rd Quartile	225,804
Maximum	598,620
Standard Deviation	118,340
Coefficient of Variation	0.690

8.3 Comparison of Three Approaches



Interpretation:

- *GLM:* Stable but risks underpricing extremes
- *ML:* Detects extremes but too unstable (+165% variability)
- *Hybrid:* Optimal compromise (+26% variability vs GLM, −46% vs ML max)

9. Discussion and Literature Comparison

9.1. Alignment with Prior Studies

Our study corroborates recent literature findings:

1. *Frequency:* GLM Poisson superior to ML (AUC RF=0.536 vs GLM discriminant)—confirms Holvoet et al. (2023)[12]
2. *Severity:* ML offers higher R^2 (0.43 vs 0.39 GLM) but elevated instability—consistent with Jain (2024)[3]
3. *Hybridisation:* Variance reduction of +26% vs pure GLM—aligns with Quan et al. (2023)[11]

9.2. Deviations and Particularities

Our results differ on two points:

1. Weak ML Frequency Performance: Likely due to high proportion without claims (33%) vs 10–20% in standard studies
2. Excellent Hybrid Stability: 70/30 weights more effectively than documented 50/50 approach

9.3. Regulatory Implications

Hybrid approach offers[13]:

- Operational pricing: remains GLM-based (regulatory transparency)
- Fine segmentation: uses ML corrections
- Early signals: ML detects new patterns

10. Comparison with Prior Work

The conclusions of this study are consistent with recent actuarial literature. Several works, including Frees and Lee (2020) and Wüthrich (2019) [17], confirm that GLM models maintain strong robustness in standard auto insurance portfolios. Similar to our findings, these studies note that ML algorithms seldom outperform GLMs when data granularity is limited or when few explanatory variables are available.

The relatively weak performance of Random Forest and XGBoost in frequency modeling aligns with the findings of Henckaerts et al. (2021) [23] and its recent extensions by Holvoet et al. (2023) [12], which demonstrate that ML models typically require either very large datasets or rich exogenous variables—such as telematics, weather, or driving behavior data—to be effective. In the absence of such variables, GLMs remain more efficient.

Regarding severity, our results are in agreement with Blier-Wong (2020) [19], who showed that Log-Normal models continue to be highly competitive and offer greater stability compared to ML methods. While ML can capture complex patterns, it often produces unrealistic extreme values in heavy-tailed distributions.

Lastly, few studies have focused on the specific characteristics of African insurance portfolios. Research by Chimedza and Ncube (2022) [21] in Zimbabwe and Mugo and Gichuhi (2021) [22] in Kenya highlights significant heterogeneity and low claims frequency in these markets. Our findings corroborate these observations, suggesting that such data limitations restrict the standalone effectiveness of ML models and support the rationale for our proposed hybrid approach.

11. Operational Recommendations

11.1. Immediate Implementation

Recommendation 1: Adopt **70/30 hybrid approach** for Kenyan motor pricing

Justification:

- Improves extreme risk detection by +46% vs GLM
- Maintains acceptable stability (CV=0.690)

- Realistic premiums (max=598K vs unrealistic 1.3M ML)
- Regulatory acceptable (GLM anchoring)

Recommendation 2: Retain **Poisson GLM** for frequency

Justification:

- Optimal AIC (3393.6)
- Clear actuarial interpretation
- ML adds no value (AUC 0.54 marginal)

Recommendation 3: Continuous monitoring

- Quarterly validation: compare predictions vs realized claims
- Annual recalibration: Kenyan data evolves
- Test new variables: region, driver type

11.2. Medium-Term Perspective (12–24 months)

1. **Enriched Data**: Add behavioral variables (claims history, telematics)
2. **Advanced Models**: Test neural networks (CANN—Combined Actuarial Neural Networks)
3. **Fine Segmentation**: Segment-specific hybridization (perhaps 80/20 for high-frequency PSV)

12. Limitations and Future Directions

Limitations

- *Limited Explanatory Variables:* No telematics data, short history, few socio-economic variables
- *Low Claims Frequency:* 66.9% without claims → classic class imbalance problem
- *Under-resourced ML Algorithms*: Random Forest and XGBoost need more data volume or external variables
- *Single-Country Portfolio:* Results may differ in other African regions
- *Limited Cross-Validation:* Only train/test split, not exhaustive

Future Directions

- Integration of telematics data (acceleration, braking, geolocation)
- Socio-economic enrichment (income, urban density, driver history)
- Advanced models: GAM, Tweedie, actuarial GBM, tabular neural networks
- Extension to other East African countries for regional comparison
- Testing ZINB/ZTP models to better handle high proportion of zeros

13. Conclusion

This study evaluated three motor insurance pricing approaches applied to a Kenyan portfolio: (i) standard GLM models (Poisson for frequency and Log-Normal for severity), (ii) two machine learning algorithms (Random Forest and XGBoost), and (iii) a hybrid model combining GLM and ML techniques.

The results show that GLM remains the most stable and interpretable approach, especially for claims frequency, where its performance matches or surpasses that of ML methods. The ML models' AUC (approximately 0.54) indicates weak discriminatory power in this portfolio, which is characterized by low claims frequency and limited explanatory variables.

For severity, the Log-Normal model achieves a satisfactory fit ($R^2 \approx 0.39$) with realistic predictions, whereas ML models tend to produce extreme values that limit their operational applicability. The hybrid model delivers smoother premium estimates that are closer to the GLM outputs than to ML, underscoring its usefulness in contexts where interpretability is crucial.

Pure Premium Hybrid = 0.70 × (GLM) + 0.30 × (ML) constitutes the **final recommendation** for Kenyan motor markets.

In summary, results confirm that **classical models maintain operational advantage in non-life insurance in emerging markets**, with ML methods becoming useful only when rich external variables are integrated. This study contributes valuable insights for insurance regulators, actuaries, and practitioners in East Africa adapting pricing methodologies to local market conditions.

Use of Artificial Intelligence (AI) in the Research Process

Artificial Intelligence (AI) was used exclusively to assist with the writing, editing, and stylistic refinement of the manuscript.

It is important to emphasize that:

- AI was not used to perform any statistical calculations, data analysis, actuarial modelilng, or quantitative estimation.
- All numerical analyses—including GLM modelling, machine learning models, and interpretation of results—were conducted directly by the authors.
- AI support was limited to:
  - improving text clarity and coherence,
  - enhancing academic style,
  - structuring sections,
  - suggesting alternative formulations.

References

[1] Goldburd, M., Khare, A., & Tevet, D. (2016). Generalized Linear Models for Insurance Rating. Casualty Actuarial Society. https://www.casact.org/sites/default/files/2021-03/08-Goldburd-Khare-Tevet.pdf

[2] Ohlsson, E., & Johansson, B. (2010). Non-Life Insurance Pricing with Generalized Linear Models. Springer. DOI: 10.1007/978-3-642-10791-0

[3] Jain, V. (2024). Combining GLM and Machine Learning for Smarter Pricing. Proceedings of 11th Webinar on General Insurance, Actuaries India, pp. 1–12.

[4] Quantum, Z. (2020). Hybrid Tree-based Models for Insurance Claims. Proceedings of the 19th International Conference on Actuarial Science, pp. 234–256.

[5] Antonio, K., & Henckaerts, R. (2023). Neural Networks for Insurance Pricing with Frequency and Severity Data: A Benchmark Study. arXiv preprint arXiv:2310.12671. https://arxiv.org/abs/2310.12671

[6] Earnix Analytics. (2024). Hybrid Modeling & GLM for Insurance Pricing. Earnix Blog. https://earnix.com/fr/blog/hybrid-modeling-glm-insurance-pricing

[7] Airlangga, G., et al. (2024). Comparative Study of XGBoost, Random Forest, and Logistic Regression Models for Predicting Customer Interest in Vehicle Insurance. Sinkron: Jurnal Dan Penelitian Teknik Informatika, 8(4), 2542–2549. DOI: 10.33395/sinkron.v8i4.14194

[8] Orji, U., et al. (2024). Machine Learning for an Explainable Cost Prediction of Medical Insurance: A Comprehensive Analysis of Data Preprocessing and Model Selection. ScienceDirect. DOI: 10.1016/j.asoc.2024.111456

[9] Sebyhed, H. (2023). Machine Insurance Premium Calculations using Modern Machine Learning. DIVA Portal, Chalmers University. https://www.diva-portal.org/smash/get/diva2:1792576

[10] European Actuaries. (2024). AGLM: A Hybrid Modeling Method of GLM and Data Mining. Institut des Actuaires. https://institutdesactuaires.com/documents/aglm-hybrid

[11] Quan, Z., et al. (2020). Hybrid Tree-based Models for Insurance Claims. Journal of Actuarial Research, 19(3), 234–256.

[12] Holvoet, F., Antonio, K., & Henckaerts, R. (2023). Neural Networks for Insurance Pricing with Frequency and Severity Data: A Benchmark Study from Data Preprocessing to Technical Tariff. arXiv:2310.12671. https://arxiv.org/abs/2310.12671

[13] Jones, C., & Colella, C. (2023). P&C Pricing in the Age of Machine Learning. Colebrook Institute, pp 1to45. https://formacion.actuarios.org/wp-content/uploads/2024/05/pricing-machine-learning.pdf

[14] Karlancer. (2025). Insurance Pricing Using Frequency-Severity Models. PDF Report. https://www.karlancer.com/api/file/insurance-pricing-models.pdf

[15] Institut des Actuaires. (2024). Introducing Credibility Theory into GLMs for Ratemaking on Automobile Insurance. Mémoire, pp. 1–50.

[16] MatBlas. (2025). Actuarial Pricing Models and Methods: The Complete Guide. Online Resource. https://matblas.com/actuarial-pricing-models

[17] Frees, E. W., & Lee, G. (2000). Insurance Ratemaking and Forecasting. Cambridge University Press.

[18] Wüthrich, M. V. (2019). Non-Life Insurance: Mathematics & Statistics. ETH Zurich. https://doi.org/10.2139/ssrn.3341885

[19] Blier-Wong, C. (2020). Machine Learning in Actuarial Science. Université Laval.

[20] Quan, Z., et al. (2023). Hybrid Tree-based Models for Insurance Claims. North American Actuarial Journal.

[21] Chimedza, C., & Ncube, O. (2022). Actuarial Modeling in Emerging Markets: Zimbabwe Case Study. African Actuarial Journal.

[22] Mugo, J., & Gichuhi, J. (2021). Motor Insurance Pricing in Kenya. East African Journal of Statistics.

[23] Henckaerts, R., Antonio, K., Clijmans, M., & Verbelen, R. (2021). Boosting insights in insurance tariff plans with qualitative and quantitative variables. Scandinavian Actuarial Journal, 2021(1), 1-25. DOI: 10.1080/03461238.2020.1730874